DHPLT: large-scale multilingual diachronic corpora and word representations for semantic change modelling

Mariia Fedorova[1], Andrey Kutuzov[1], Khonzoda Umarova[2]
mariiaf@ifi.uio.no, andreku@ifi.uio.no, ku47@cornell.edu
(equal contribution, authors sorted in alphabetical order)
[1]University of Oslo (Norway), [2]Cornell University (USA)

Abstract

In this resource paper, we present DHPLT, an open collection of diachronic corpora in 41 diverse languages. DHPLT is based on the web-crawled HPLT datasets; we use web crawl timestamps as the approximate signal of document creation time. The collection covers three time periods: 2011-2015, 2020-2021 and 2024-present (1 million documents per time period for each language). We additionally provide pre-computed word type and token embeddings and lexical substitutions for our chosen target words, while at the same time leaving it open for the other researchers to come up with their own target words using the same datasets.

DHPLT aims at filling in the current lack of multilingual diachronic corpora for semantic change modelling (beyond a dozen of high-resource languages). It opens the way for a variety of new experimental setups in this field.

1 Introduction

Computational data-driven diachronic semantic change modelling (tracing meaning shifts over time) naturally requires diachronic corpora: that is, texts annotated with their creation date. Once such datasets are obtained, one can compare the usage of target words (or any other linguistic phenomena) in different time periods, using any preferred change modelling method.

However, diachronic corpora of appropriate size and quality are not easy to find, especially permissively licensed. Most of the current lexical semantic change detection (LSCD) projects operate on the same small set of high-resource languages. For example, the seminal SemEval 2020 Task 1 on LSCD (Schlechtweg et al., 2020) was limited to English, German, Latin and Swedish. Later, LSCD benchmarks based on diachronic corpora for Italian (Basile et al., 2019), Greek (Perrone et al., 2019), Russian (Kutuzov and Pivovarova, 2021), Norwegian (Kutuzov et al., 2022a), Spanish (Zamora-Reina et al., 2022), Chinese (Chen et al., 2023), Japanese (Ling et al., 2023), Finnish (Fedorova et al., 2024b), and Slovene (Pranjić et al., 2024) were presented, but not more than that (not for all languages the corpora themselves are publicly available). The field has to experiment with at most a dozen of languages, with Indo-European family strongly over-represented. This limits the scope of LSCD research, especially on *multilingual* semantic change effects.

To fill in this gap, we release **DHPLT** ('Diachronic HPLT'): a set of standardized diachronic corpora for 41 languages of 12 different language families. Each language is represented with three time-dependent subsets, containing 1 million documents each. These documents are extracted from the web-crawled datasets by the HPLT project, specifically HPLT v3.0 (Oepen et al., 2025): thus, they are basically cleaned and filtered web pages in the target language. We use crawling timestamps as the signal for time period separation (see below).

In addition, we define a set of potentially interesting 'target words' for each language. For the DHPLT occurrences of these words, we produce a variety of semantic representations (static `word2vec` embeddings, token embeddings, lexical substitutions). This allows practitioners to start experimenting with multilingual LSCD immediately, without spending compute on re-creating these representations. At the same time, the availability of the original texts makes it possible to come up with other target word sets. **All the resources described in this paper are available at `https://data.hplt-project.org/three/diachronic/`, sorted by language.**

2 Diachronic corpora out of HPLT

In prior work, diachronic resources for LSCD mostly were produced from existing historical corpora manually created by linguists: newspaper

The Proceedings for the 6th International Workshop on Computational Approaches to Language Change (LChange'26), pages 87–96

archives, releases by national libraries, etc. Unfortunately, such resources are nearly non-existent for the majority of world's languages: at least in anything resembling a standardized form. In an ideal world, fragmented national efforts in historical corpora creation could be unified and merged into a multilingual diachronic resource. But the amount of work required for such a project is well beyond the scope of this paper or any research group we are aware of. That's why we instead suggest to rely on the Internet as the source of diachronic data.

World Wide Web contains hundreds of billions of documents in all existing languages and of varying quality (many documents consist of SEO keywords, machine-generated slop or price lists). At least two initiatives are currently crawling the WWW and saving representative slices of its state: Common Crawl[1] (CC) and Internet Archive[2] (IA). The HPLT project (Burchell et al., 2025) processes these web crawls by conducting language identification, deduplication, cleaning, etc, to produce language-specific corpora of competitive quality.[3] Importantly, all its datasets are published under the Creative Commons CC0 license. HPLT v3.0[4] is the specific data release we are using.

HPLT provides a lot of clean documents, but to create a diachronic corpus, we need to know the *date* when the document was created. It is impossible to label all the HPLT documents with the creation date manually. Sometimes, web pages do contain the date of their publication either in plain text or in some structured form. But this data is not reliable: it is perfectly possible for a web document to be published in 2024, but contain text created in 2001. Also, creating parsing rules for all sorts of HTML creation date labels would be an immense effort - with no guarantee that the result will fully reflect the diversity of the Web. Thus, we instead rely on a different time signal: web crawling time stamps. All the HPLT documents can be traced back to specific web crawls and they inherit the 'timestamp': that is, the exact date and time when a given web page was downloaded and saved.

Admittedly, these timestamps do not directly map to the creation date of the document: again, it is absolutely possible for CC or IA in 2024 to download a web page created in 2001. But the timestamps do provide an 'upper boundary' of the creation date: if some text was crawled in 2015, there is no way for it to have been created later than 2015. Web crawl timestamps allow us to create diachronic datasets of a sort slightly different from 'traditional' diachronic corpora. Here, subsets for periods 1, 2 and 3 contain documents created *no later than 1, 2 or 3* respectively. Importantly, the subset 3 can still contain documents created in the earlier time periods: but not vice versa. For sure, this is less precise than manually labelled historical corpora: but we believe this still can be an important source of diachronic text data.

We aim at a diachronic dataset with more than two time periods, since this makes it possible to conduct research in long-term multipoint dynamics of semantic change (Kutuzov and Pivovarova, 2021). We also would like our time periods to be as comparable as possible in terms of the amount of data, and to be separated by at least some 'gaps', since this makes it easier to detect semantic change (Giulianelli et al., 2022). In order to choose the exact temporal spans, we analyse the distribution of documents in the HPLT v3.0 datasets by the year of crawling. Figure 1 in the Appendix shows these numbers for English and Georgian as an example. Our main observations are that 1) 2011 is the earliest crawl year, and the number of documents remains relatively low until 2017; 2) much more documents were crawled in 2020 and after, with peaks in 2020 and 2024 (the latest crawl year).[5]

Based on these observations, we come up with the following three time periods, each 2-4 years long, and with gaps of at least two years: 2011-2015 ('*early time period*'), 2020-2021 ('*Covid time period*'), 2024 ('*most recent crawls*'). The three-time-period structure for the DHPLT is useful for studying and capturing linguistic innovation or the onset of semantic change at different points in time. With the crawl timestamps being the 'upper bound' on the document creation time, we can, for instance, observe the rise of the concept of 'remote work' in 2020-2021 and then look at its journey in 2024.

Note that our time bins are far from being the only possible choice. We consider them to be a sensible way of temporally splitting the existing HPLT data, but depending on the objective, other splits can make more sense. All the documents

[1] https://commoncrawl.org/

[2] https://archive.org/

[3] Another project from which one could extract diachronic web corpora is FineWeb 2 (Penedo et al., 2025).

[4] https://hplt-project.org/datasets/v3.0

[5] Interestingly, 2013 is a rather rare crawl year in HPLT v3.0: many languages have no documents crawled in 2013.

in our datasets are accompanied with full timestamps, so anyone can produce their own subsets of DHPLT: for example, more fine-grained. It is also possible to use our open source code[6] to reproduce DHPLT from the original HPLT data with any desired changes.

We produce three subsets of the HPLT v3.0 datasets containing documents crawled during the time periods above. But first, we need to choose what *languages* DHPLT will contain.

2.1 Language selection

The original HPLT v3.0 datasets feature 198 languages, which is way too many for our purposes. Our selection of languages for DHPLT is based on the following criteria:

1. Language must have at least 0.5 million documents in each of the time periods above: smaller languages do not provide sufficient amount of data and also are more error-prone with regards to language identification.

2. There should exist a corresponding HPLT v3.0 `T5` monolingual encoder-decoder language model[7] (Oepen et al., 2025): we use these models to generate token embeddings in 4.

As a result, we come up with a set of 41 languages. Table 1 in the Appendix lists them along with their ISO codes (augmented with the writing system code) and the corresponding number of documents in each of the three time periods.

2.2 Data extraction pipeline

For each of the languages, we construct three time-specific corpora by randomly sampling 1 million documents from the HPLT v3.0 dataset corresponding to the given language and time period. Where less than 1 million documents are available, we only sample 0.5 million. The resulting diachronic corpora are published as `zstd`-compressed JSONL files, following the HPLT format, with the total size $\approx$ 170 GB (for comparison, the full size of HPLT v3.0 is 50 TB), and $\approx$ 59 billion words.

These diachronic corpora can already be used for multilingual LSCD research. However, we also provide more 'refined' data for practitioners: namely, semantic representations (which can be computationally expensive for academic researchers to produce) for pre-defined sets of 'target words'. They are described in the next sections. One can think about them as an *example* of what sorts of experimental setups are possible with DHPLT.

[6] https://github.com/ltgoslo/scdisc_hplt

[7] https://hf.co/collections/HPLT/hplt-30-t5-models

3 Target word selection

For each language we select a subset of the vocabulary – target words – representations of which would be part of DHPLT. Our primary objective in selecting target words is to narrow down the full corpus vocabulary while keeping as many words that would be of interest to lexical semantic change researchers as possible.

Starting from the T5 model vocabulary corresponding to each given language, we filter out word pieces and infrequent tokens, leaving only words that appear as nouns, verbs or adjectives and are written in the language's main script. Please refer to Appendix B for full details of our target word selection process.

Our selection pipeline yields a set of target words for each DHPLT language, with the average size of $\approx$ 18600 (HPLT T5 models' vocabulary size is 32768). Figure 2 shows the distribution of target word counts across languages.

Target lemmas We additionally lemmatize each of the resulting target words. For instance, in English distinct target tokens 'thread', 'Thread', and 'threads' share one common lemma **'thread'**. We later use lemmas to merge word representations into more linguistically-informed groupings.

4 Target word representations

Once the language-specific target words are defined, we produce a number of different semantic representations for their occurrences in the DHPLT corpora. These representations can be directly used by LSCD practitioners to evaluate or train semantic change models on the three DHPLT time periods.

4.1 Contextualized word embeddings

Contextualized token embeddings are widely utilized in lexical semantic-change research (Periti and Tahmasebi, 2024; Umarova et al., 2025, inter alia). They can serve both as direct representations that are later averaged into prototypical embeddings (Periti and Montanelli, 2024) and as a basis for constructing clusters corresponding to different 'sense nodules' (Martinc et al., 2020; Kutuzov et al., 2022b). For our DHPLT dataset, we obtain encoder embeddings for 1000 randomly sampled occurrences per target word from HPLT v3.0 T5

monolingual models and the XLM-R model (Conneau et al., 2020); we additionally produce encoder embeddings from 100 randomply sampled occurrences per target word from HPLT v3.0 GPT-BERT (Charpentier and Samuel, 2024) monolingual models[8] (Oepen et al., 2025).

4.2 Lexical substitutes

In addition to language model embeddings, we also consider lexical substitutes as a different kind of contextualized representations. Substitutes-based semantic change quantification methods were shown to do well at both LSCD benchmarks (Card, 2023; Periti et al., 2024) and downstream tasks like semantic change discovery (Umarova et al., 2025).

While it is possible to perform masked language modelling with T5 models, the results are not suitable for lexical substitutions generation, since these models were pre-trained with the span masking objective and tend to predict longer sequences rather than single lexemes. Examples can be found in Appendix C. For this reason, we use the HPLT v3.0 GPT-BERT models in a way similar to Card (2023) and Umarova et al. (2025) to represent 100 randomly sampled occurrences of each target word via top-15 substitutes. More details can be found in Appendix D. We also release XLM-R lexical substitutions. The number of target words for which we provide XLM-R embeddings and substitutions is limited by the size of the intersections between XLM-R and HPLT T5 tokenizer vocabularies. HPLT v3.0 GPT-BERT models use exactly the same tokenizers as the corresponding HPLT T5 models, so this issue is not relevant for them.

4.3 Word type embeddings

Although approaches based on *contextualized token embeddings* (as the ones described above) are the 'daily drivers' of modern LSCD researchers (Periti and Montanelli, 2024), we also publish *static type embedding* models trained on the DHPLT corpora. Static word embedding (SWE) models yield one vector representation per word type, as opposed to 'a representation for each word occurrence' from contextualized models. They were the LSCD mainstream until around 2021-2022 (Schlechtweg et al., 2020) and are still often used because of their simplicity and relatively modest compute requirements, both for training and for inference.

We train a SWE model for each language/time period combination using the SGNS architecture, also known as `word2vec` (Mikolov et al., 2013). For simplicity, we mostly use training hyperparameters from Aida and Bollegala (2025): window size 10, 5 epochs, 5 negative samples. Our embedding size is set to 300, and the model vocabularies are limited to 50000 most frequent words. Before training, the DHPLT documents are filtered to remove punctuation, as well as leading and trailing tabs and whitespaces.

Finally, for each language, the vector spaces of the models trained on time periods 1 (2011-2015) and 2 (2020-2021) are *aligned* to the vector space of the model trained on time period 3 (2024-), so as to make it possible to directly compute similarities between word embeddings in different models. We do this with the standard Procrustes alignment technique (Hamilton et al., 2016).

4.4 Frequency counts

Finally, we also publish frequency counts of each target word across the three DHPLT time periods. Changes in frequency of word usage coupled with lemma information are some of the very first indicators of changes in the word usage. These counts can also be used to control for frequency effects when quantifying semantic change (Card, 2023) and for planning compute usage (e.g. when generating lexical substitutions, half of the time is used for finding samples of the 100 least frequent target words, following the Zipf's law (Powers, 1998)).

5 Sanity check

To demonstrate the utility of the DHPLT diachronic corpora, we look at the SWEs of the English word *'AI'* ('artificial intelligence'). Table 4 shows how the semantics of this term drifted. Back in the beginning of 2010s, it was associated almost exclusively with 'AI characters' in video games. A decade later, in 2020-2021, AI starts to be associated with 'chatbots' and machine learning, but it is still very much about robots, unmanned vehicles and Internet of Things. Only in the very last period 3 (2024-) we see the much too familiar landscape of LLMs, ChatGPT and 'generative AI'. Interestingly, this trajectory is clear even despite the fact that (as noted above), our corpus for period 2 is bound to contain some documents created during the period 1, and the corpus for period 3 surely contains some documents from both 1 and 2.

[8] https://hf.co/collections/HPLT/hplt-30-gpt-bert-models

We observe a similar pattern (Table 5) when looking at the SWEs of the Spanish equivalent of the word: 'IA' ('inteligencia artificial'). In 2010s, the word often appears in the context of gaming: 'jugabilidad', 'PS', 'BETA', etc. Then in 2020-2021, we start seeing words like 'algoritmos' and 'tecnología(s)' among closest semantic neighbours. These words follow more semantically-similar English-words such as 'AI', 'artificial', 'learning', etc, which are likely to still appear in the Spanish DHPLT corpora as part of names and titles. Finally, in 2024 'IA' starts being associated with 'generativa' and 'ChatGPT'. Very similar trends are found in our SWEs trained on Russian DHPLT documents (Table 6).

For T5 encoder embeddings, we calculate average pairwise distance between different time period representations (Kutuzov et al., 2022b) for the English lemmas 'ai', 'remote', 'legislative', and 'jurisdiction'. The exact scores are to be found in the Appendix E. The change of 'ai' semantics is the largest and corresponds to the aforementioned SWE findings, while the changes of 'legislative', and 'jurisdiction', which are terms from the conservative legal domain, are the smallest. The degree of 'remote' change is somewhere in between and is the largest between 2011-2015 and 2020-2021 ('Covid') periods, when this word began to refer specifically to remote work rather than other contexts[9]. These observations hold for Spanish DHPLT corpora as well (Table 3).

6 Conclusion

We present DHPLT ('Diachronic HPLT'): an open collection of large-scale diachronic corpora in 41 languages of 12 different language families. It is based on the web-crawled HPLT v3.0 datasets (Oepen et al., 2025), using web crawl timestamp as the temporal signal. The collection covers three time periods: 2011-2015, 2020-2021 and 2024. We augment DHPLT with pre-computed token-level semantic representations for language-specific sets of target words, to make it easier for practitioners to start experimenting with our corpora. Finally, we provide aligned static (type-based) word embedding models for each language and time period.

DHPLT (partially) addresses the lack of multilingual diachronic corpora in the LSCD field. We hope it will help making the landscape of historical language change modelling more rich and diverse. It should be especially relevant for studies in semantic change discovery.

Limitations

The main limitation of DHPLT is the source of temporal signal: that is, web crawl timestamps. As described above, timestamp X on a document does not guarantee that the text in this document was not created in the time periods $< X$ (earlier than X). It guarantees only that the text was not created in the time periods $> X$ (later than X). This difference compared to traditional diachronic corpora should be kept in mind when working with DHPLT.

Another limitation is that we provide only some of the possible types of semantic representations, and only for selected sets of target word, not for *all* the words in each of our languages. This is inevitable, given compute and storage space constraints. DHPLT allows practitioners to come up with their own sets of target words, or even conducts semantic change discovery experiments on the entire corpus. Our representations were obtained from models that do not incorporate any temporality, thus, e.g. masked language modeling predictions for 2011-2015 can include proper names that in fact first emerged in later periods. For future work, one might employ approach from (Fittschen et al., 2025) and pre-train models on texts from corresponding periods only.

Finally, in this paper we only introduce the DHPLT dataset; we leave conducting full-scale semantic change discovery on it for future work.

Acknowledgments

The computations were performed on resources provided by Sigma2 - the National Infrastructure for High-Performance Computing and Data Storage in Norway. This work was also in part supported by a gift from Google. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Google.

References

Taichi Aida and Danushka Bollegala. 2025. SCDTour: Embedding axis ordering and merging for interpretable semantic change detection. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14775–14785, Suzhou, China. Association for Computational Linguistics.

[9] https://languages.oup.com/word-of-the-year/2020/

Pierpaolo Basile, Giovanni Semeraro, and Annalina Caputo. 2019. Kronos-it: a dataset for the Italian semantic change detection task. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, pages 423–428, Bari, Italy. CEUR Workshop Proceedings.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 2025. An expanded massive multilingual dataset for high-performance language technologies (HPLT). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

Dallas Card. 2023. Substitution-based Semantic Change Detection using Contextual Embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 590–602, Toronto, Canada. Association for Computational Linguistics.

Lucas Georges Gabriel Charpentier and David Samuel. 2024. GPT or BERT: why not both? In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.

Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023. ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 93–99, Singapore. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mariia Fedorova, Andrey Kutuzov, and Yves Scherrer. 2024a. Definition generation for lexical semantic change detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5712–5724, Bangkok, Thailand. Association for Computational Linguistics.

Mariia Fedorova, Timothee Mickus, Niko Partanen, Janine Siewert, Elena Spaziani, and Andrey Kutuzov. 2024b. AXOLOTL'24 shared task on multilingual explainable semantic change modeling. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 72–91, Bangkok, Thailand. Association for Computational Linguistics.

Elisabeth Fittschen, Sabrina Li, Tom Lippincott, Leshem Choshen, and Craig Messner. 2025. Pretraining language models for diachronic linguistic change discovery. *Preprint*, arXiv:2504.05523.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics.

Mario Giulianelli, Andrey Kutuzov, and Lidia Pivovarova. 2022. Do not fire the linguist: Grammatical profiles help language models detect semantic change. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 54–67, Dublin, Ireland. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Sinan Kurtyigit, Maike Park, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Lexical semantic change discovery. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6985–6998, Online. Association for Computational Linguistics.

Andrey Kutuzov and Lidia Pivovarova. 2021. Three-part diachronic semantic change dataset for Russian. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 7–13, Online. Association for Computational Linguistics.

Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittemann. 2022a. NorDiaChange: Diachronic semantic change dataset for Norwegian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France. European Language Resources Association.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022b. Contextualized embeddings for semantic change detection: Lessons learned. In *Northern European Journal of Language Technology, Volume 8*.

Zhidong Ling, Taichi Aida, Teruaki Oka, and Mamoru Komachi. 2023. Construction of evaluation dataset for Japanese lexical semantic change detection. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 125–136, Hong Kong, China. Association for Computational Linguistics.

Nikola Ljubešić, Luka Terčon, and Kaja Dobrovoljc. 2024. CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. In *Conference on Language Technologies and Digital Humanities (JT-DH-2024)*, Ljubljana, Slovenia. Institute of Contemporary History.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, Barry Haddow, Jan Hajič, Jindřich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkonen, Bhavitvya Malik, Vladislav Mikhailov, and 13 others. 2025. HPLT 3.0: Very large-scale multilingual resources for LLM and MT. Mono- and bilingual data, multilingual evaluation, and pre-trained models. *Preprint*, arXiv:2511.01066.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. FineWeb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Francesco Periti, Pierluigi Cassotti, Haim Dubossarsky, and Nina Tahmasebi. 2024. Analyzing semantic change through lexical replacements. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4495–4510, Bangkok, Thailand. Association for Computational Linguistics.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Computing Surveys*, 56(11):1–38.

Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.

Valerio Perrone, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2019. GASC: Genre-aware semantic change for Ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 56–66, Florence, Italy. Association for Computational Linguistics.

David M. W. Powers. 1998. Applications and explanations of Zipf's law. In *New Methods in Language Processing and Computational Natural Language Learning*.

Marko Pranjić, Kaja Dobrovoljc, Senja Pollak, and Matej Martinc. 2024. Tracking semantic change in Slovene: A novel dataset and optimal transport-based distance. *arXiv preprint arXiv:2402.16596*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Khonzoda Umarova, Lillian Lee, and Laerdon Kim. 2025. Current semantic-change quantification methods struggle with discovery in the wild. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35342–35355, Suzhou, China. Association for Computational Linguistics.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.

A DHPLT datasets

Figure 1 shows the number of documents from different crawls for English and Georgian languages in HPLT v3.0.

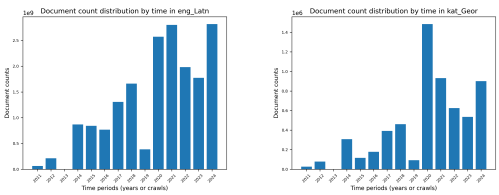

Figure 1: Number of documents per crawl year in the HPLT v3.0 datasets: English (left) and Georgian (right).

DHPLT files contain one document per line, with the following data fields:

- `id`: unique document identifier, can be used to link back to the original HPLT dataset,
- `ts`: timestamp, the exact date and time when the document was crawled from the Web,
- `text`: the actual document body, split into 'segments' (in most cases equal to paragraphs) with line break symbols,
- `doc_scores`: a list of integer 'quality scores' assigned to each segment of the document; to produce these scores, HPLT project employs heuristics from `Web Docs Scorer` (WDS).[10]

Table 1 lists all the DHPLT languages and their statistics.

B DHPLT target words

Below we provide technical details on selection of the target words.

For a given language L, we start from the vocabulary V_{T5_L} of the corresponding T5 model from the HPLT v3 T5 model collection. These models were pre-trained on documents in specific languages from the original HPLT v3.0 dataset. We assume that words which don't appear as their own token in the corresponding T5 vocabulary are not frequent enough, so we omit them.

Further, we exclude tokens from V_{T5_L} that are word pieces or non-words. To do this, we first employ a heuristic for identifying full words. For most languages after removing punctuation we split the documents from diachronic corpora by whitespace, and count occurrences of such 'full words'. For languages where splitting by whitespace doesn't make sense, we use specific splitters. For Japanese, we employ `fugashi`[11] library that utilizes dictionaries for tokenization. For simplified Chinese, we segment words in text using `jieba`[12] library. Finally, for Thai, we go with word tokenization via `pythainlp`[13] library.

Next, we count occurrences of such 'full words' across the diachronic corpora and filter out infrequent terms using a minimum frequency threshold. Thus, we only keep a token from V_{T5_L} if it appears as a 'full word' at least 10 times in each of the three time periods (i.e., at least 30 times across all diachronic corpora in that language). Note that we ignore case when counting frequencies: e.g., if 'operation' and 'OPERATION' each occurs 5 times in the corpus, we keep both.

Further, following Kurtyigit et al. (2021), we also limit target words to only nouns, verbs, or adjectives. We use Stanza (Qi et al., 2020) part-of-speech taggers for all languages except Macedonian, for which we use `classla` (Ljubešić et al., 2024). In cases where the tagger identifies a token as a proper noun but its lower-cased version is tagged as a noun, the 'NOUN' tag takes precedence. Additionally, with the exception of Chinese, Japanese and Korean, we remove single character

Language	ISO Code	Family	1	2	3
Albanian	als_Latn	Indo-European	0.5M	1M	1M
Arabic	arb_Arab	Afro-Asiatic	1M	1M	1M
Bosnian	bos_Latn	Indo-European	1M	1M	1M
Bulgarian	bul_Cyrl	Indo-European	1M	1M	1M
Catalan	cat_Latn	Indo-European	1M	1M	1M
Czech	ces_Latn	Indo-European	1M	1M	1M
Chinese	cmn_Hans	Sino-Tibetan	1M	1M	1M
Danish	dan_Latn	Indo-European	1M	1M	1M
German	deu_Latn	Indo-European	1M	1M	1M
Estonian	ekk_Latn	Uralic	0.5M	1M	1M
Greek	ell_Grek	Indo-European	1M	1M	1M
English	eng_Latn	Indo-European	1M	1M	1M
Finnish	fin_Latn	Uralic	1M	1M	1M
French	fra_Latn	Indo-European	1M	1M	1M
Hebrew	heb_Hebr	Afro-Asiatic	1M	1M	1M
Croatian	hrv_Latn	Indo-European	1M	1M	1M
Hungarian	hun_Latn	Uralic	1M	1M	1M
Armenian	hye_Armn	Indo-European	0.5M	1M	0.5M
Indonesian	ind_Latn	Austronesian	1M	1M	1M
Italian	ita_Latn	Indo-European	1M	1M	1M
Japanese	jpn_Jpan	Japanese	1M	1M	1M
Georgian	kat_Geor	Kartvelian	0.5M	1M	0.5M
Korean	kor_Hang	Korean	1M	1M	1M
Lithuanian	lit_Latn	Indo-European	1M	1M	1M
Latvian	lvs_Latn	Indo-European	0.5M	1M	1M
Macedonian	mkd_Cyrl	Indo-European	0.5M	1M	1M
Dutch	nld_Latn	Indo-European	1M	1M	1M
Norwegian	nob_Latn	Indo-European	1M	1M	1M
Polish	pol_Latn	Indo-European	1M	1M	1M
Portuguese	por_Latn	Indo-European	1M	1M	1M
Romanian	ron_Latn	Indo-European	1M	1M	1M
Russian	rus_Cyrl	Indo-European	1M	1M	1M
Slovak	slk_Latn	Indo-European	1M	1M	1M
Slovenian	slv_Latn	Indo-European	1M	1M	1M
Spanish	spa_Latn	Indo-European	1M	1M	1M
Swedish	swe_Latn	Indo-European	1M	1M	1M
Tamil	tam_Taml	Dravidian	0.5M	1M	1M
Thai	tha_Thai	Tai-Kadai	1M	1M	1M
Turkish	tur_Latn	Altaic	1M	1M	1M
Ukrainian	ukr_Cyrl	Indo-European	1M	1M	1M
Vietnamese	vie_Latn	Austro-Asiatic	1M	1M	1M

Table 1: DHPLT languages, writing systems, language families and historical period sizes (in millions of documents).

[10] https://github.com/pablop16n/web-docs-scorer
[11] https://github.com/polm/fugashi
[12] https://github.com/fxsjy/jieba
[13] https://github.com/PyThaiNLP/pythainlp

tokens in all languages.

Finally, we ensure that words in the target set are written in this language's main script. Following the approach similar to `alphabet-detector`[14], we use the Unicode names of characters in a target word candidate to verify its script. For instance, for English, we check that each character in each word from the target set is from Latin Unicode blocks, while for Japanese, we check that characters are either Hiragana, Katakana, or Kanji. We exclude words in which at least one character doesn't belong to the expected script.

Figure 2 shows the distribution of target word counts across languages in DHPLT.

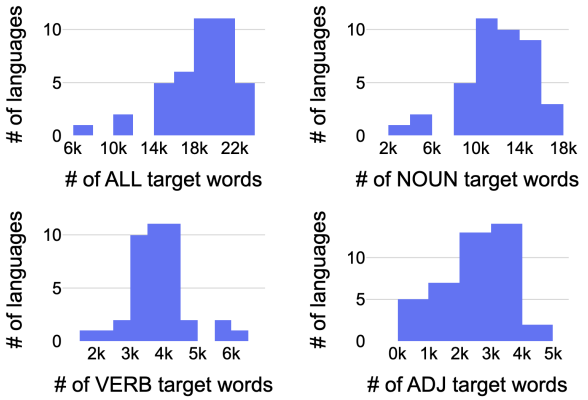

Figure 2: Number of target words across 41 languages for all target words (top left), target words that are nouns (top right), verbs (bottom left), and adjectives (bottom right).

C T5 substitutions

We use the same input format in our experiments, as shown in T5 model cards[15]. The reasons for not using T5 to generate lexical substitutions are :

- both encoder and decoder parts are required, duplicating the compute usage
- while it is technically possible to obtain not only the one most probable, but also top-k predictions with `beam search` (Freitag and Al-Onaizan, 2017), in practice the difference between predictions is too vague to make them useful. For example, the input *I remember we [MASK_1] it but still - do we use the decoder part because the final hidden states of the encoder part are not mapped to vocabulary logits?* (with the word 'discussed' behind the mask) yields the predictions 'used the decoder part to do', 'used the decoder part for', and 'used the decoder part before'. Such small differences in generated representations are rather a problem for LSCD than an advantage, as shown in Fedorova et al. (2024a).
- predictions do not necessarily represent the semantics of the target word and tend to repeat other terms from the sentence, which is also observed on the aforementioned example
- longer inputs from real-world HPLT documents yield even longer predictions, for example, *How do you know the apples you are using for hard cider are ripe? Maybe, you would [MASK_1] me to define ripe. Is ripe defined by the ideal time to harvest an apple, to eat an apple, or to press an apple. We could even consider the question of ripeness for cooking apples. In my ... Continue reading When are apples ripe?* yields *'say that the apples you are using for hard cider are ripe. But I don't think that is the right way for'*

D BERT representations

HPLT 3.0 T5s and HPLT 3.0 GPT-BERTs use the same tokenizer vocabulary for each language, so each target word has representations produced by both models.

E Examples

E.1 Sanity check for T5 embeddings

Tables 2 and 3 show change degrees of the English ('ai', 'remote', 'legislative', 'jurisdiction') and Spanish ('ia', 'remoto', 'legislativo', 'jurisdicción') words, according to the average pairwise distance, APD method (Fedorova et al., 2024a), on their respective T5 token embeddings.

Period pairs	'ai'	'remote'	'legislative'	'jurisdiction'
1 to 2	0.5533	0.4586	0.4117	0.4495
1 to 3	**0.5646**	0.4619	0.4141	0.4497
2 to 3	0.48	0.4548	0.4191	0.4351

Table 2: Average pairwise distances for several English target words calculated on T5 encoder embeddings.

[14] https://github.com/EliFinkelshteyn/alphabet-detector

[15] https://huggingface.co/collections/HPLT/hplt-30-t5-models

Period pairs	'ia'	'remoto'	'legislativo'	'jurisdicción'
1 to 2	0.5733	0.5104	0.4031	0.4470
1 to 3	0.5763	0.4955	0.3925	0.4438
2 to 3	**0.5810**	0.4821	0.3979	0.4423

Table 3: Average pairwise distances for several Spanish target words calculated on T5 encoder embeddings.

E.2 Sanity check of HPLT 3.0 GPT-BERT substitutions

We perform manual analysis of the same 4 English words as in Section 5. The observations obtained with SWE models still hold: in 2011-2015, the words predicted as substitutions for 'ai' are either non-technical, or related to games or cars. In 2020-2021, a wider range of technologies is mentioned, including 'IoT', 'NLP', 'robotics', 'animation', etc. Also a lot of terms reflecting the social influence of AI emerge: 'cybersecurity', 'humanity', 'innovation', names of states and companies. Finally, in 2024, the trend of discussing social consequences continues: 'elite', 'censorship', 'communism', 'scammers', 'capitalism'; much less technical terms and much more human-related ones are observed. There are also mentions of spheres which traditionally were human-dominated but has become automated recently: 'art', 'healthcare' etc. Surprisingly, we don't observe many 'LLM'-related terms among GPT-BERT's predictions, but rather a shift from the optimistic perception of AI to the pessimistic one.

In 2011-2015, 'remote' is associated with networks and being spatially (geographically) distant. In 2020-2021, 'virtual' frequently occurs. In 2024, the associations show a techno-optimistic pattern similar to that of 'AI' in 2020-2021: positive job-related adjectives ('skilled', 'flexible', 'professional'), wider range of technologies and spheres ('satellite' and 'healthcare' emerge). We also see terms related to society: state names, 'climate', 'rural'.

Substitutions of 'jurisdiction' and 'legislative' bring no surprises, being related to law throughout all three time periods.

To conclude, representations obtained from contextualized models are sensitive to particular contexts at prediction, and thus capture more fine-grained semantic nuances than SWE models.

E.3 Sanity check for SWEs

Table 4 shows the semantic trajectory of the English word 'AI' in the DHPLT time periods, according to our static word embedding models (SWEs). Similarly, Table 5 shows the trajectory for 'IA' that stands for 'inteligencia artificial' in Spanish across the three DHPLT time periods. Table 6 does the same for the Russian abbreviation 'ИИ' ('AI') (the model trained on the first time period does have this word in its vocabulary).

1: 2011-2015	2: 2020-2021	3: 2024-
multiplayer	chatbots	generative
NPCs	IoT	AI's
RPG	robotics	GenAI
animations	RPA	ChatGPT
FPS	intelligence	LLMs

Table 4: Top 5 nearest neighbours (by cosine similarity) of the English term 'AI' in DHPLT static word embedding models by time periods. Case is ignored.

1: 2011-2015	2: 2020-2021	3: 2024-
BETA	AI	generativa
PS	artificial	artificial
AI	algoritmos	AI
jugabilidad	learning	inteligencia
artificial	inteligencia	ChatGPT

Table 5: Top 5 nearest neighbours (by cosine similarity) of the Spanish term 'IA' in DHPLT static word embedding models by time periods. Case is ignored.

2: 2020-2021	3: 2024-
интеллект (intellect)	интеллект (intellect)
AI	нейросети (neural networks)
роботов (robots)	ChatGPT
блокчейн (blockchain)	AI
алгоритмы (algorithms)	искусственный (artificial)

Table 6: Top 5 nearest neighbours (by cosine similarity) of the Russian term 'ИИ' ('AI') in DHPLT static word embedding models by time periods. Case is ignored. The 2011-2015 model does not have the word in its vocabulary (because of low frequency in this time period).